Interactive visualization of gene regulatory networks with associated gene expression time series data

Westenberg, Michel A.; Hijum, Sacha A.F.T. van; Lulko, Andrzej T.; Kuipers, Oscar P.; Roerdink, Jos B.T.M.

Published in:
VISUALIZATION IN MEDICINE AND LIFE SCIENCES

Interactive Visualization of Gene Regulatory Networks with Associated Gene Expression Time Series Data

Michel A. Westenberg[1], Sacha A. F. T. van Hijum[2], Andrzej T. Lulko[2], Oscar P. Kuipers[2], and Jos B. T. M. Roerdink[1]

[1] Institute for Mathematics and Computing Science, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands m.a.westenberg@rug.nl, j.b.t.m.roerdink@rug.nl

[2] Department of Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, P.O. Box 14, 7950 AA Haren, The Netherlands s.a.f.t.van.hijum@rug.nl, a.t.lulko@rug.nl, o.p.kuipers@rug.nl

Summary. We present GENeVis, an application to visualize gene expression time series data in a gene regulatory network context. This is a network of regulator proteins that regulate the expression of their respective target genes. The networks are represented as graphs, in which the nodes represent genes, and the edges represent interactions between a gene and its targets. GENeVis adds features that are currently lacking in existing tools, such as mapping of expression value and corresponding p-value (or other statistic) to a single visual attribute, multiple time point visualization, and visual comparison of multiple time series in one view. Various interaction mechanisms, such as panning, zooming, regulator and target highlighting, data selection, and tooltips support data analysis and exploration. Subnetworks can be studied in detail in a separate view that shows the network context, expression data plots, and tables containing the raw expression data. We present a case study, in which gene expression time series data acquired in-house are analyzed by a biological expert using GENeVis. The case study shows that the application fills the gap between present biological interpretation of time series experiments, performed on a gene-by-gene basis, and analysis of global classes of genes whose expression is regulated by regulator proteins.

1 Introduction

The unraveling of interactions between components of living cells is an important aspect of systems biology. The interaction networks are very complex, since interactions take place not only at genomic, proteomic, and metabolomic levels, but also between these levels. We are establishing a software framework

that is able to visualize such networks, and which offers interactive exploration to a researcher [BBO+06]. As part of this effort, we have developed an application for visualization of gene regulatory networks.

Gene regulatory networks can be represented by graphs, in which nodes represent genes, and edges represent interactions between a gene product (a regulator protein) and its target genes. The nodes have several attributes, such as position on the chromosome, a Gene Ontology classification [The00], and in our case, they also have gene expression attributes for multiple time points acquired during distinctive phases of growth together with p-values indicating statistical significance or other statistical data. Gene expression is measured in terms of the amount of messenger RNA (mRNA) produced after transcription of the gene. A number of tools have been proposed that visualize gene networks and overlay gene expression data on the network [BST03, SMO+03, HMWD04, HMW+05, BSRG06]. These tools overlay the expression value of one time point on a node, often as the node color, and do not always map the associated statistical data to a visual representation or one that is easy to interpret. However, proper analysis and interpretation is not possible without statistical confidence information. A further problem is that none of the existing tools allows a researcher to overlay multiple time points and associated statistical confidence on nodes. However, simultaneous visualization of multiple time points would make discovery of trends and outliers much easier. Similarly, it is also not possible to compare multiple time series with each other in a single view.

In this paper, we present GENeVis (*G*ene *E*xpression and *Ne*twork *Vis*ualization), an application that allows a researcher to simultaneously visualize gene regulatory networks and gene expression time series data. Our application extends on concepts introduced in previous work in gene regulatory network visualization, and it adds features that are currently lacking in existing tools, such as mapping of expression value and corresponding p-value (or other statistic) to a single visual attribute, multiple time point visualization, and visual comparison of multiple time series in one view. We have used GENeVis to analyze time series data of the bacterium *Bacillus subtilis*, acquired in-house [LBKK07], in its regulatory network context, acquired from DBTBS (DataBase of Transcriptional Regulation in *Bacillus subtilis*) [MNON04].

The organization of this paper is as follows. We briefly discuss previous work in Section 2. We then describe the design of our application in detail (Section 3), and present the case study (Section 4). Conclusions are drawn in Section 5.

2 Previous Work

There exist a large number of tools that allow visualization of general graphs, see Herman et al. [HMM00] for an overview. In the bioinformatics field, a

number of tools have emerged more or less independently due to specific requirements from the biological community (see [BBO+06] for an overview).

Osprey [BST03] was one of the first biological interaction network visualization tools. Genes are colored by their biological process as defined by the Gene Ontology [The00]. Osprey cannot overlay time series expression data on the interaction network. Cytoscape [SMO+03] is a popular data analysis tool, which does support visualization of gene expression attributes. Statistical attributes associated with the expression data can be mapped to visual styles of nodes, such as color, shape, size, border width, and border color. A main shortcoming of Cytoscape is that, for time series, it can show only the expression value of a single time point. Though extensible through a plug-in mechanism, the design of Cytoscape makes it hard to incorporate alternative node visualization methods.

VisANT [HMWD04, HMW+05] is a tool for biological network analysis and visualization. It focusses strongly on the analysis of network of various types, such as protein-protein interaction networks, gene transcription networks, metabolic pathways, and interconnections between these. Network analysis is performed by calculating topological statistics and features, or querying a server-side database for functional information. The strength of VisANT is the integration of multiple network data sources for a large number of species, which is a hard problem due to naming convention issues between data sets. For analyzing and visualizing gene expression data, VisANT is not suitable, since it provides no support for loading such data.

BiologicalNetworks [BSRG06] is a tool with a strong focus on data integration and analysis. It supports visualization of time series gene expression data in matrix form and by plots, but only allows the user to overlay one time point at a time on an interaction network.

Recently, Saraiya et al. [SLN05] performed a user performance study for various graph and time series visualizations. This study is of particular interest, since the test case reflects tasks that are performed commonly in bioinformatics pathway analysis. Metabolic pathways are also represented as graphs, forming a kind of flow chart of the chemical reactions and genes involved in, for instance, some biological process. The data for the test case consisted of time series gene expression data for 10 time points and a 50-node directed graph. Their source of data remains unclear. The study involved four visualization approaches: (i) single attribute (showing one time point at a time) and single view (show only the graph), (ii) single attribute and multiple views (show the graph and a parallel coordinate linked view), (iii) multiple attribute (show all time points simultaneously) and single view, and (iv) multiple attribute and multiple views. Statistical data associated with the expression data were not included in the experiment. It was found that overlaying a single attribute at a time works well for analyzing graphs at particular time points, and for search tasks that require topological information. Showing multiple attributes simultaneously reduces user performance for such tasks. On the other hand, multiple attribute visualizations result in better performance in outlier search

tasks, and also in node comparison tasks between two time points. The conclusion was that visualization design should be task specific.

Despite the existence of tools for gene expression analysis in a regulatory network context, we believe and demonstrate in our case study that biologists would benefit from a richer and more interactive visualization environment to analyze their time series data. Existing tools overlay only expression values of one time point at a time on a node, and usually have a poor visual mapping of the statistical properties of the data. Furthermore, none of the tools supports comparison of multiple time series, i.e., in which each gene is associated with multiple time points obtained from multiple time series. GENeVis provides a solution to these issues.

3 Visualization Design

We will now present the design of GENeVis, and describe our choices regarding graph layout, expression mapping, and possibilities for interaction and data exploration.

3.1 Graph Layout

The layout of the network is computed by a force-directed algorithm, in which the edges act as springs, and the nodes repel each other [BETT99]. This layout algorithm produces satisfactory layouts for the type of networks in our application, since the grouping of nodes corresponds quite well with the biological concept of a regulon (a collection of genes under regulation by the same protein). The nodes (genes) are drawn as boxes and the edges (interactions) as lines between the nodes. A regulator protein can inhibit or activate its target, which is represented graphically at the target end of an edge by a bar or an arrow, respectively. It can also be the case that the interaction type is unknown, in which case the edge is not decorated. An example is shown in Fig. 1, in which a part of the regulatory network of *Bacillus subtilis* is drawn. The gene boxes are annotated with the gene name, and we can see, for example, that *gerE* inhibits *cotA* and *spoIVCB*, activates *cotB*, *cotC* and a number of other genes, and that it also interacts with *cotD* and *sigK* in an unknown way.

3.2 Gene Expression Mapping

Exploration of gene regulatory networks is often based on time series gene expression data, where the growth of an organism or tissue type is followed in time. For bacteria, one could take measurements during early, middle, late exponential, and stationary growth phases, resulting in a time series containing four points. This amount is small enough to allow visualization of all time

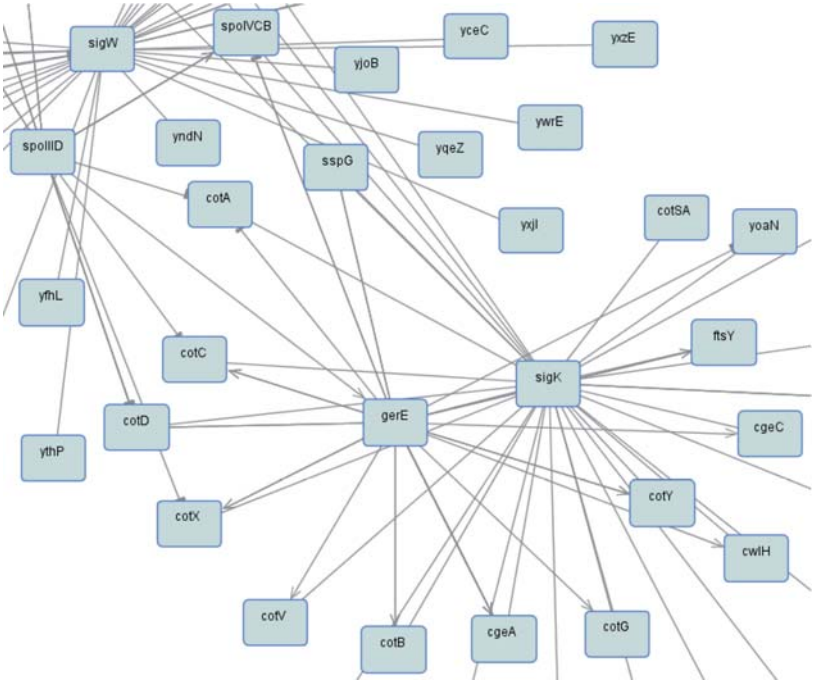
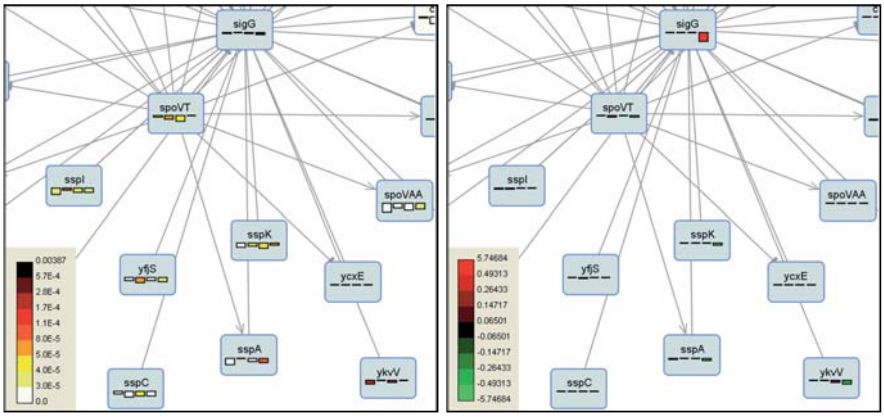


Fig. 1. Part of the regulatory network of *B. subtilis*. Gene boxes are annotated with their respective gene names. Graph edges represent gene interactions, where bars and arrows at the target ends represent inhibition and activation, respectively. Undecorated edges are used when the type of interaction is not known.

points simultaneously, and we map each time point to a colored expression box drawn inside the gene box. As a time series experiment is very time consuming and also expensive, a maximum number of 50 time points would be a realistic limit. Our approach can be used also for these larger time series. A larger number of time points will usually correspond to higher resolution in time, which can be mapped in a visually intuitive way by reducing the width of the expression boxes.

Gene expression values can either be absolute levels of expression or ratios between a test condition and a reference condition. To each expression value, a statistical value is associated, which expresses the reliability of the measurement. Commonly, the coefficient of variation is used in the case of expression levels and a p-value (indicated by p in the remainder of this paper) is used in the case of expression ratios. The reliability value is used to scale the height of an expression box: the more reliable, the higher the box. Expression levels are mapped to colors that range from white to black via yellow and red. Expression ratios are mapped to colors that range from green to red via black. The use of these colormaps is standard practice in the bioinformatics field. We divide the expression data range into a number of quantiles, and assign



(a) (b)

Fig. 2. Visualization of expression levels (a) and expression ratios (b) for four time points overlaid on a part of the gene regulatory network of *B. subtilis*.

each quantile a color from the colormap. By inspecting the expression value range corresponding to each quantile, a user can obtain some insight in the statistical distribution of the data. We selected a colored rectangular glyph as a graphical representation of the expression data, since color and size are perceptually easy to separate and interpret independently [War04].

Figure 2(a) and Fig. 2(b) show a visualization of expression levels and expression ratios, respectively. In order to demonstrate the visual mapping, an arbitrary part of the regulatory network of *B. subtilis* is shown. More details about the gene expression data can be found in Section 4. At all times, the user can refer to the color legend for the statistical distribution of the data. Note that the expression values are mapped in a nonlinear way to colors through the use of quantiles. Figure 2(a), for instance, shows that, at time point three, *spoVT* has a low expression (yellow color), and that this is measured reliably. The expression box would be square for a high confidence, and it reduces to a line for a very low confidence. In Fig. 2(b), we can see that *sigG* is strongly up-regulated (its expression level is high in comparison with the expression level of the reference) in time point four, and that this is the only significant change in this part of the network.

3.3 Interaction

The user can interact with the visualization by simple mouse operations. Panning and zooming are performed by dragging the mouse with the left or right button down, respectively. A right mouse button click in the background causes an automatic pan and zoom, such that the entire network fits within the display bounds. A small overview display containing a view of the whole network and a semitransparent rectangle corresponding to the area visible in

the main display helps the user to navigate through the network, see Fig. 6. The rectangle in the overview display supports user interaction, and it can be dragged to pan the display.

3.4 Exploration

Even though a force-directed layout algorithm produces acceptable layouts, it is sometimes difficult to understand the network structure in dense areas with highly interconnected nodes. Therefore, we have implemented a mechanism that highlights a gene and its direct targets when the mouse hovers over the gene. Highlighting increases the line widths of the gene boxes and the edges, and it colors the edges according to the type of regulation. Activation maps to the color green, inhibition to the color red, and unknown interactions map to a shade of grey. Figure 3 shows an example when the user hovers the mouse over the gene *gerE*.

Tooltips are used to display additional information about a specific gene. This information includes the gene locus (the position of the gene on the chromosome), the gene name and possibly synonyms, gene function, and a

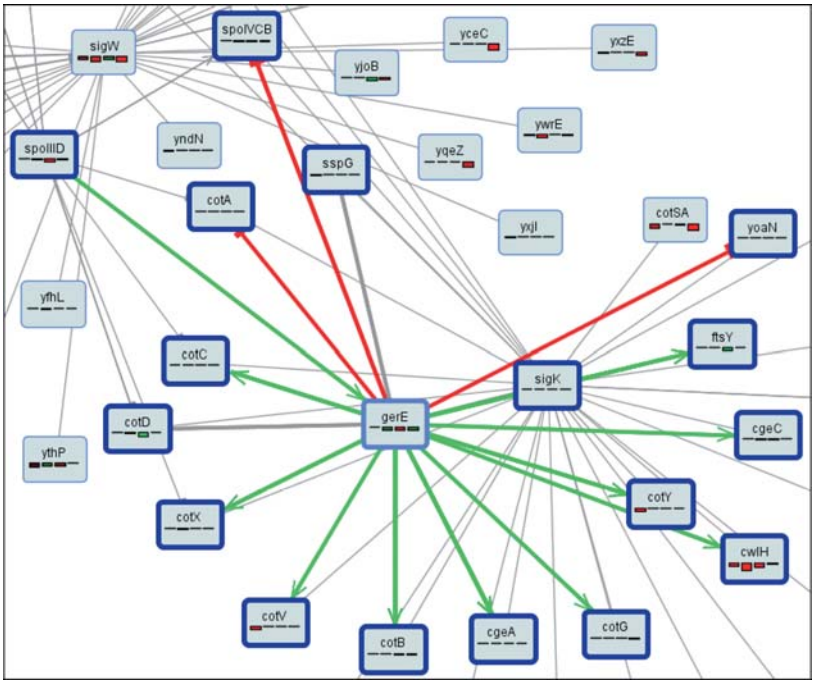


Fig. 3. Neighbor highlighting assists the user in understanding network structure. The interaction type is mapped to a color: red for inhibition, green for activation, and grey for other cases.

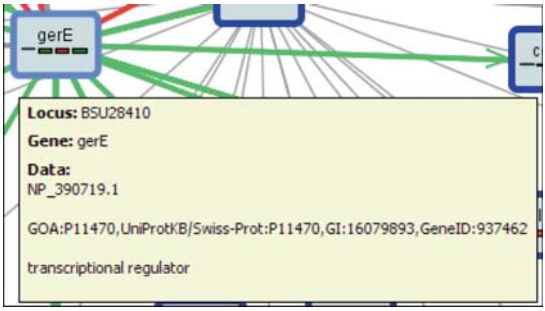


Fig. 4. Tooltips display additional information about a gene.

list of gene identifiers for other databases (e.g., Gene Ontology Annotation, Universal Protein Resource). The tooltips appear also by hovering the mouse over a gene box (i.e., in combination with the highlighting effect described previously) see Fig. 4 for an example.

GENeVis also supports keyword search to aid the user in finding a specific gene. When the gene exists in the network, the display automatically pans, such that the found gene is centered. As an additional visual cue, the gene box is enlarged slightly.

Our application also supports single time point analysis. The user can choose a time point, and construct a filter that selects genes for which the measurements are statistically significant, and for which the expression levels or ratios fall within a certain range. The filter parameters are specified by sliders and radio buttons (see Fig. 6 top left). Interaction with the sliders and buttons provides immediate visual feedback: the background of each selected gene box is filled with a color corresponding to its expression value; the background of a gene box that is not selected is set to a standard color that is not in the color map. In this way, a user can quickly spot particular behavior at specific time points, and answer questions such as "which genes are up-regulated during early growth?".

To study the interaction between a specific gene and its direct targets, it is possible to select the corresponding subnetwork by right-clicking a node. This action opens a new window that contains the subnetwork, plots of the expression values, and a table containing the raw gene expression data. The table lists expression values and corresponding statistical data. Both plot and table show only expression profiles than contain at least one time point for which the significance falls within the significance range limits. This range is controlled by the significance slider that is also used for single time point analysis, as explained above. This view allows a biologist to consider the expression data qualitatively in the network visualization, but also quantitatively by inspecting the raw data or the corresponding plots. To assist the user in maintaining a mental map of the complete network, the layout of the subnetwork is not changed. An example is shown in Fig. 5, which contains the gene *ccpA* and

its targets. The expression plot shows the ratio measures at all time points of only significant data ($p < 0.00001$). The gene box backgrounds are colored according to time point 1.

For a large regulator, the plot can become cluttered, even when only significant data are shown. Therefore, the user can also add or remove the expression profile of a gene by left-clicking its gene box in the network visualization. The profile and corresponding raw data will then be added to (or removed from) the plot and the table below the plot, respectively.

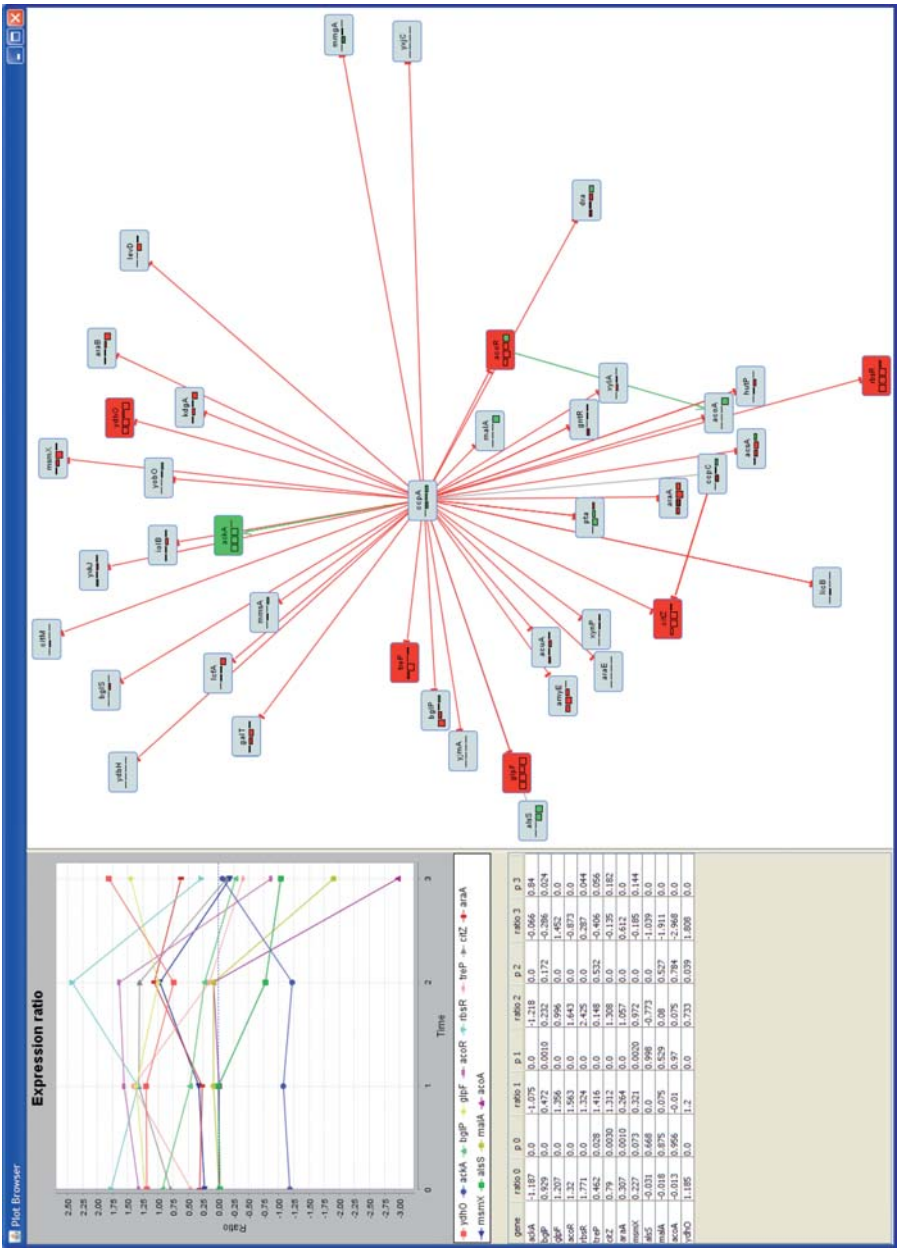
3.5 Implementation

Our application was built with use of the Prefuse library [HCL05], which is an open source Java toolkit for interactive information visualization. The toolkit provides basic data structures for storing graphs and node and edge attributes, supports many layout algorithms, and has a flexible rendering mechanism. Expression profile plotting was implemented with JFreeChart[Gil06], a free Java chart library, distributed under the LGPL.

Prefuse provides basic functionality to calculate a layout, and to color data based on some attribute. It uses Java2D to draw the graphs in a display that supports panning and zooming. The display also provides a handle to tooltips. We have implemented extensions of the standard edge and node renderers (those perform the actual drawing on the display). The edge renderer was modified such that it decorates the edge with an arrow or a bar depending on the interaction type. The node renderer was extended such that in addition to a text label containing the gene name, it also draws the gene expression boxes. Other components, such as the overview display, were created by combining modules of basic functionality already present in Prefuse.

4 Case Study

GENeVis has been used to further explore a short time series DNA microarray dataset described by Lulko and coworkers [LBKK07]. In this study, the global mRNA levels (thus gene transcription or gene expression) at four distinct stages of growth of the bacterium *B. subtilis* strain 168 and the same strain containing a gene deletion are compared. These four growth stages were sampled to obtain a view of the changes in gene expression during growth of this bacterium. The four time points sampled ranged from (i) the early exponential phase (the onset of fast cell growth), (ii) mid-exponential phase (fast cell growth), (iii) end-exponential phase (nutrients start slightly limiting the growth), and (iv) the stationary phase of growth (no growth of cells and start of cell death). The *B. subtilis* strain with a gene deletion has its *ccpA* gene disabled, and it is therefore called a *ccpA* deletion mutant. This comparison was performed by DNA microarrays.



gene	ratio 0	p 0	ratio 1	p 1	ratio 2	p 2	ratio 3	p 3
araA	-1.187	0.0	-1.075	0.0	-1.218	0.0	-0.066	0.84
bglP	0.929	0.0	0.472	0.0010	0.232	0.172	+0.386	0.024
glpF	1.207	0.0	1.356	0.0	0.996	0.0	1.452	0.0
acoR	1.32	0.0	1.563	0.0	1.643	0.0	+0.873	0.0
rbsR	1.771	0.0	1.324	0.0	+2.425	0.0	0.287	0.044
treP	0.462	0.028	1.416	0.0	0.148	0.532	+0.406	0.056
cdZ	0.79	0.0030	1.312	0.0	1.308	0.0	+0.135	0.182
araA	0.307	0.0010	0.264	0.0	1.057	0.0	0.612	0.0
msmX	0.227	0.073	0.321	0.0020	0.972	0.0	+0.185	0.144
alsS	+0.031	0.668	0.0	0.998	+0.773	0.0	+1.039	0.0
malA	+0.018	0.875	0.075	0.529	0.08	0.527	+1.911	0.0
acoA	+0.013	0.966	+0.01	0.97	0.075	0.784	+2.968	0.0
ybhO	1.185	0.0	1.2	0.0	0.733	0.039	1.808	0.0

Fig. 5. Visualization of the subnetwork consisting of *ccpA* and its direct targets. The plot and table show the expression ratios for all time points of only the significant data ($p < 0.00001$). Data from time point 1 were used to color the gene box backgrounds.

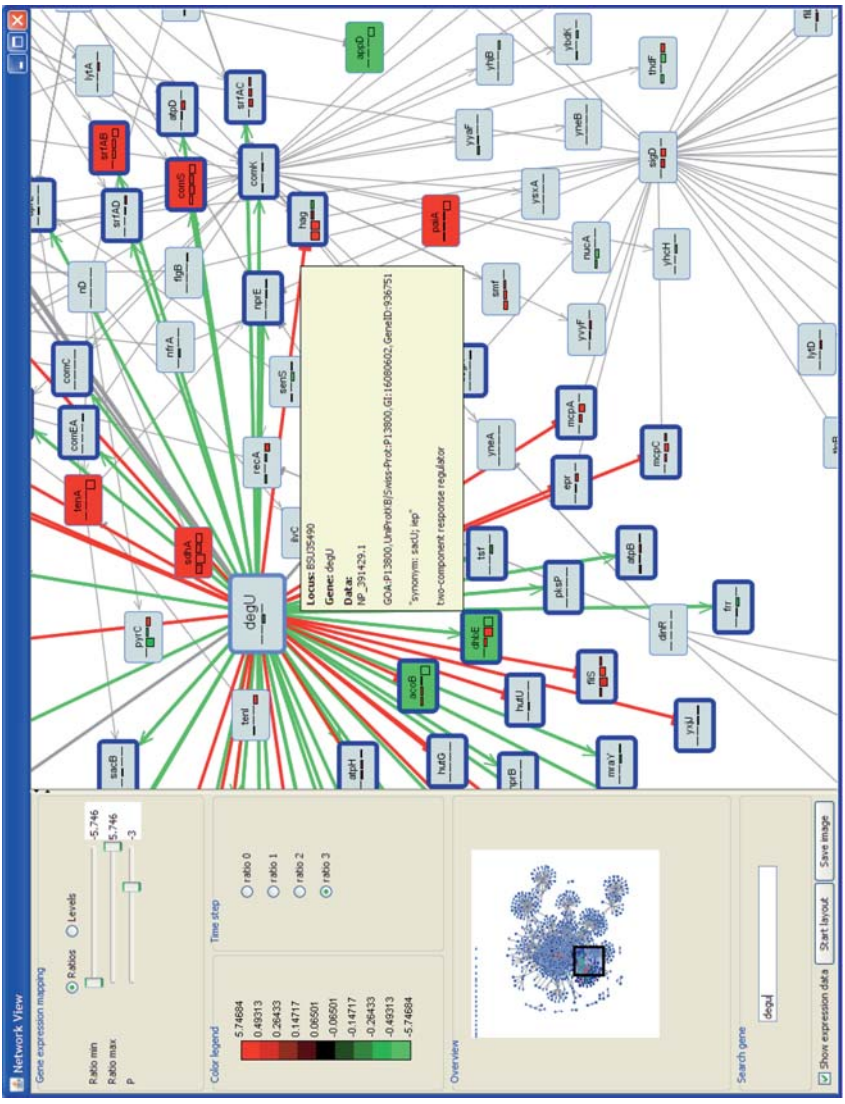


Fig. 6. Screen shot of GENeVis. The user can switch between expression level and expression ratio visualization with the radio buttons in the top left corner. The sliders together with the time point radio buttons can be used to select a subset of genes that have expression data within the ranges set by the sliders and buttons. The background of the selected genes is then colored according to their expression value. The color legend is shown to the left of the time point selection buttons. An overview display is shown to help the user in maintaining a view of the network context. The square indicates the area of the network visible in the main display, and it can be dragged to pan the view in the main display. The search box can be used to find a specific gene, and when found, the display automatically pans such that the gene found is centered in the view.

With DNA microarrays, global mRNA levels (which indicate the "activity" of genes) are determined by comparing a reference (e.g. the *B. subtilis* wild-type) to a test (e.g. the deletion mutant) condition. All genes of *B. subtilis* are present on the DNA microarray which allows monitoring the expression of these genes during the four growth-stages sampled. After quantification and normalization of the signals of the DNA microarray, the researcher is left with signals for each gene and for each condition (4 time points and 2 samples). These signals indicate the relative gene expressions. A large problem for an experimentalist is to identify relevant biological phenomena from all these measurements (in this case over 100,000 signals of over 4000 genes). Often, a ratio is used to relate the changes in expression between two conditions. This ratio is calculated by dividing the signal of the test condition (in this study the mutant) by the signal of the reference condition (in this study wild-type) for each gene and for each time point.

The CcpA protein (for which the *ccpA* gene codes)[1] is a master transcriptional regulator (a protein that drives the expression of target genes) involved in governing carbon catabolite repression in many so-called Gram-positive bacteria [SH00, WL03]. As is shown in the study of Lulko and coworkers and other studies, the inactivation of the *ccpA* gene has broad implications for a large number of key cellular processes [LBKK07, BHL$^+$03, LCB$^+$05, MSM$^+$01]. From a biologists' point of view, it is crucial to have a possibility to oversee the global changes caused by any kind of interference or indirect effects of, in this case, the deletion of a regulator gene. Furthermore, after global changes have been identified, a biologist needs to delve into the behavior of specific genes as we will demonstrate below. Current research is switching from single time point analysis to monitoring the changes of gene expressions over time in time series DNA microarray experiments. As we will show, following gene expression in time allows a richer description of the direct and indirect effects of, e.g., a gene deletion. Analysis of time series data was originally performed in a gene-by-gene approach of the most differentially expressed genes (genes whose expression was most notably changed) involving literature search and mining of information available at public repositories such as PubMed (`www.pubmed.org`). Furthermore, the global effects of the *ccpA* deletion on the known regulators and metabolic pathways were studied by Lulko and coworkers by using FIVA [BBvH$^+$07]. This tool presents an overview of the key cellular processes affected, but it does not allow visual or manual identification of groups of genes exhibiting correlated behavior within these processes. For instance, the software will indicate a regulon (a collection of genes under regulation by the same protein) affected, but not which members of the regulon. Therefore, after identification of affected key cellular processes, the experimentalist has to mine the data manually. This makes the

[1] The biological convention is that gene names are written in italics with the first letter in lower case; the corresponding protein for which the gene codes is written in roman with the first letter in upper case.

investigation of the (indirect) effects of the *ccpA* gene deletion on parts of regulons in these four time points very difficult and time consuming.

The application presented in this paper allows the projection of time series expression data derived from DNA microarrays on a gene interaction network (the network of regulator proteins which drive the expression of target genes, which in turn can also encode regulator proteins). This not only facilitates the biological interpretation of the overall direct effects of the *ccpA* gene deletion but also offers an opportunity to focus on some indirect responses caused by the disruption of this transcriptional regulator. Mining the time-series data starts with a visual exploration of the whole interaction network. By using GENeVis with an overview of the network and cycling through the four time points it immediately becomes apparent that the impact of the *ccpA* mutation dynamically develops and intensifies during growth of *B. subtilis* cells. These overviews are shown in Fig. 7 for time points 0 and 1 and in Fig. 8 for time points 3 and 4. The significant genes ($p < 0.001$) are colored according to expression ratio (calculated by the gene expression levels of the *ccpA* mutant over those of the wild-type). A red color indicates a higher expression level in the mutant strain compared to the wild-type strain, whereas a green color indicates a lower expression level. These images clearly show that (i) the number of colored genes increases strongly from time point 2 to 3, and (ii) about half of the genes in the dashed box are differentially expressed (green or red).

The regulon in the dashed box (SigB) is shown enlarged in Fig. 9. The expression of the SigB regulon in time point 3 is a prominent indication of the dynamics just described. This regulon is involved in the response to harmful environmental conditions, such as heat, osmotic, acid, or alkaline shock. In previous studies, only minor effects during the exponential phases of growth have been reported for SigB. However, the visualization of ratio-based data of the SigB regulon genes as a function of time allows a spectacular view on the reprogramming of the SigB-dependent gene expression at later growth stages (the transition from the late exponential to stationary phase of growth). From Fig. 9, three distinctive gene clusters can be identified; (i) a few genes (*csbA*, *csbX*, *yfhK*) whose expression levels in the wild-type strain, compared to the mutant strain, were lower in one of the time points during exponential growth (red-colored expression boxes inside the gene boxes for time points 0 to 2); (ii) a few genes (*dps*, *spoVG*, *yqgZ*, *yvyD*) for which there is a clear switch in the expression profile between the late exponential and the stationary phase time points (the expression box color changes from red to green for time point 3); and (iii) a larger number of genes (colored by green backgrounds) whose expression is strongly increased in the wild-type strain in the stationary phase compared to the three exponential growth phase time points during which transcripts levels remained essentially unaffected. The latter cluster is particularly interesting since it explicitly reveals that the SigB regulon is recruited stronger during the late growth stages of the wild-type strain than the *ccpA* deletion strain. Measurements of glucose concentration (a major

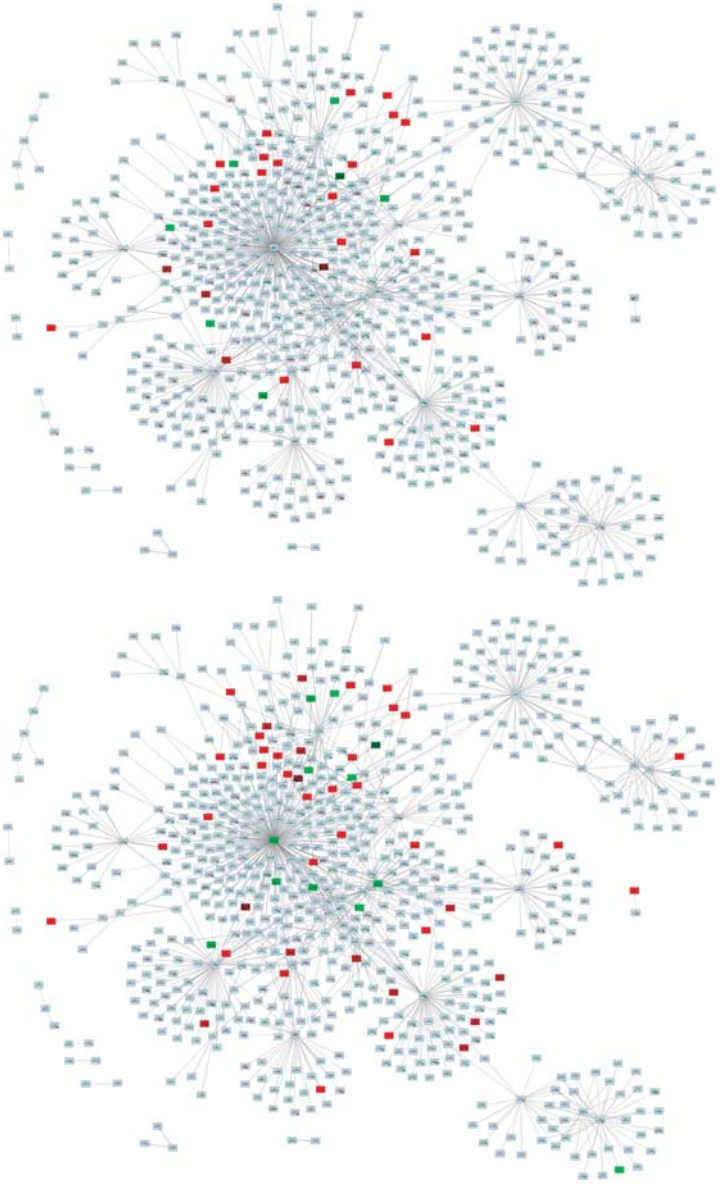


Fig. 7. A gene regulatory network of *B. subtilis* superimposed with DNA microarray ratio data of *B. subtilis* wild-type over its *ccpA* deletion mutant. The time series consists of four time points corresponding to different phases of growth. Top image: the early exponential growth phase (the onset of fast cell growth). Bottom image: mid-exponential growth phase (fast cell growth). Significantly expressed genes ($p < 0.001$) are colored to their expression ratio; others have a neutral background color. Continued in Fig. 8.

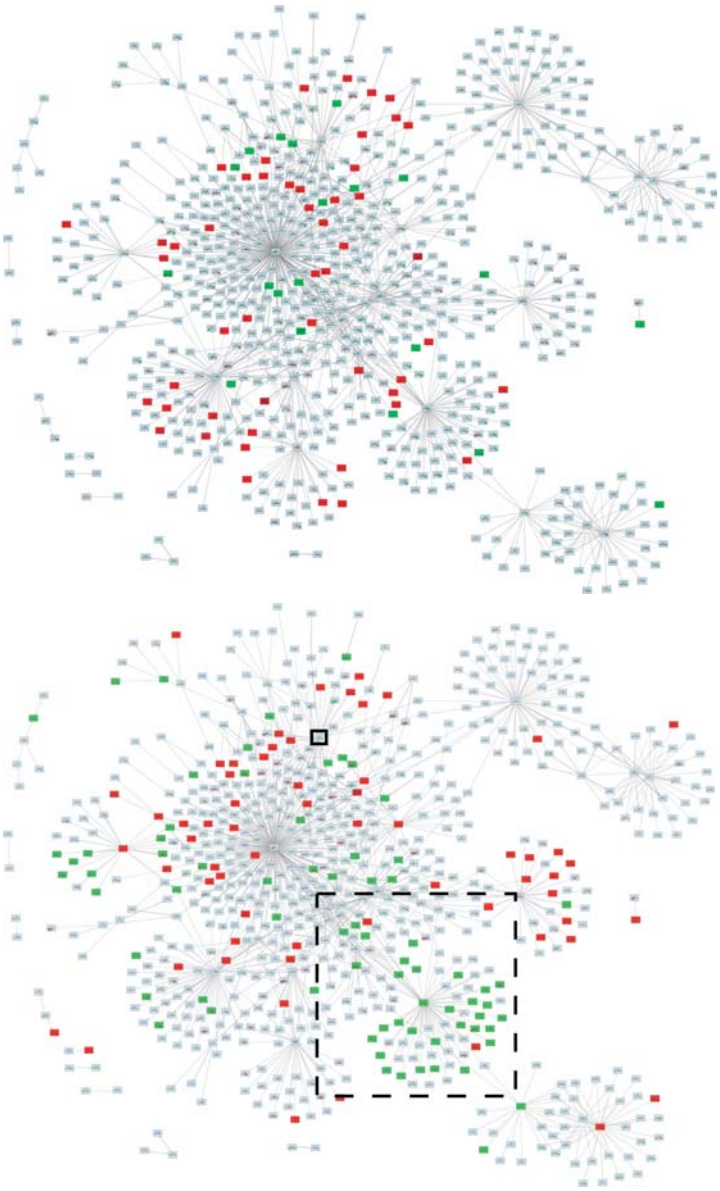


Fig. 8. Continued from Fig. 7. Top image: end-exponential growth phase (nutrients start slightly limiting the growth). Bottom image: stationary growth phase (no growth of cells and start of cell death). The *ccpA* gene is indicated by a solid black rectangle. The dashed rectangle indicates the SigB regulon, which is shown enlarged in Fig. 9.

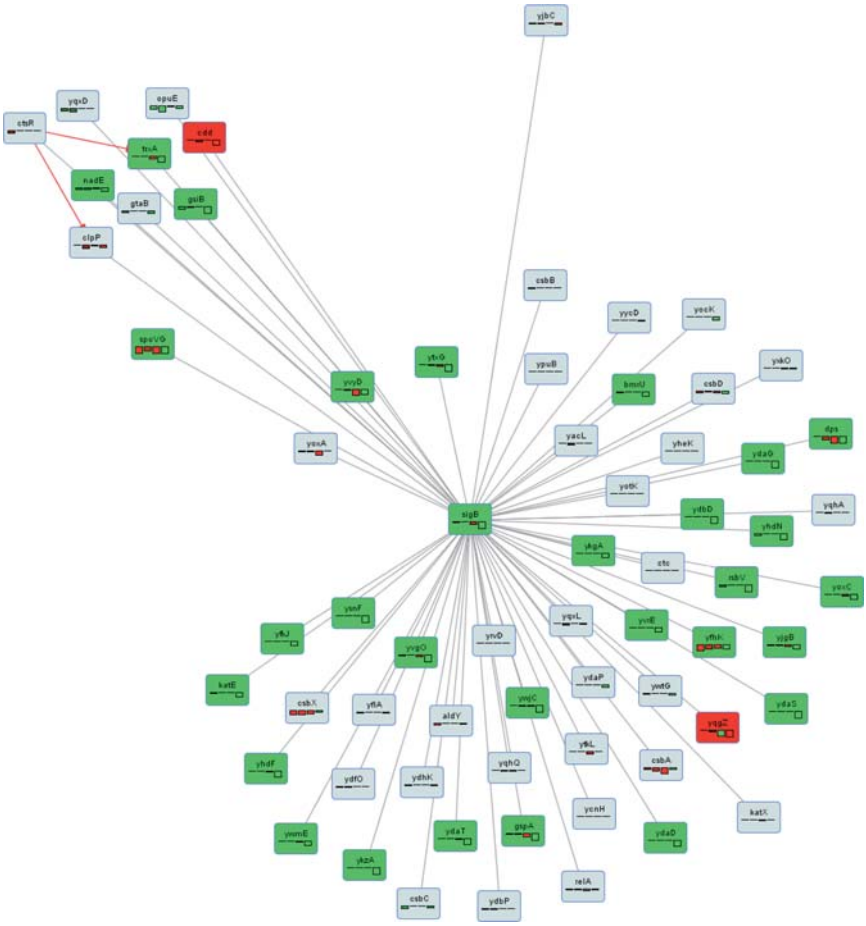


Fig. 9. The SigB regulon with gene coloring based on expression ratios ($p < 0.001$) in the stationary growth phase (time point 3). A red color indicates a higher expression level in the mutant strain compared to the wild-type strain, whereas a green color indicates a lower expression level.

source of energy for growth of *B. subtilis*; results not shown) demonstrated that glucose is completely consumed by the wild-type strain in the stationary phase, while it is still present in the culture of the mutant. This could explain the apparent induction of several members of this general stress/starvation regulon. This energy depletion due to lack of glucose apparently is the signal for the wild-type strain to adapt to the upcoming stress conditions (nutrient limitation). The induction of genes whose products counteract the oxidative stress (stress to bacterial cells caused by oxygen radicals) conditions (*dps*, *katE*, *nadE*, *trxA*) is just one of the examples indicative of bacterial adaptation to survive deteriorating environmental circumstances.

The analysis we have performed in this section could not have been done with the tools described in Section 2. Some of these tools, Cytoscape, for example, would be able to produce overviews of the entire network for the individual time points. The main problem is to visualize only significant expression ratios, which currently can only be done by preprocessing the data and removing the insignificant entries. This is very inconvenient, since the user may wish to modify the p-value range during this exploration phase. In the second part of the analysis, we have studied the behavior of the SigB regulon over time. This was only possible by looking at all time points simultaneously, which is only supported by GENeVis and not by any of the tools discussed in Section 2.

5 Conclusions

We have presented GENeVis, an application to visualize gene expression time series data in a gene regulatory network context. GENeVis adds features that are currently lacking in existing tools, such as mapping of expression value and corresponding p-value (or other statistic measures) to a single visual attribute, multiple time point visualization, and visual comparison of multiple time series. Various interaction mechanisms, such as panning, zooming, regulator and target highlighting, data selection, tooltips, and support for subnetwork analysis facilitate data analysis and exploration.

We have presented a case study, in which gene expression time series data acquired in-house have been analyzed by an end-user. Our case study has revealed that GENeVis clearly fills the gap in the present gene-by-gene biological interpretation of time series experiments and global regulon analysis. This goal is achieved by (i) allowing the biologist an overview of the gene regulatory network with mapped gene expressions as a function of time to quickly identify biologically relevant changes in parts of the network, and (ii) delve into detail by visual identification of the partitioning of members of regulons as a function of time. We have shown that the combination of single and multiple time point visualization and filtering based on statistical significance is very powerful and helpful for a biologist. The analysis presented in the case study could not have been performed otherwise, i.e., by existing visualization tools.

In the case study, we have used GENeVis to visualize the relatively small network of *B. subtilis* (772 nodes and 1179 edges). For networks of that size, a force-directed layout algorithm produces satisfactory layouts. However, this type of layout algorithm does not scale very well for larger networks. In future work, therefore, we plan to investigate other layout algorithms, such as the multi-level algorithm based on topological features [AMA07]. Another promising algorithm is the grid layout algorithm that takes both connection structure and biological function associated to the genes into account [LK05].

References

[AMA07] D. Archambault, T. Munzner, and D. Auber. TopoLayout: Multi-level graph layout by topological features. *IEEE Trans. Visualization and Computer Graphics*, 13(2):305–317, 2007.

[BBO+06] D. W. J. Bosman, E.-J. Blom, P. J. Ogao, O. P. Kuipers, and J. B. T. M. Roerdink. MOVE: A multi-level ontology-based visualization and exploration framework for genomic networks. *In Silico Biology*, 7:0004, 2006.

[BBvH+07] E. J. Blom, D. W. J. Bosman, S. A. F. T. van Hijum, R. Breitling, L. Tijsma, R. Silvis, J. B. T. M. Roerdink, and O. P. Kuipers. FIVA: Functional information viewer and analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics*, page btl658, 2007.

[BETT99] G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, New Jersey, 1999.

[BHL+03] H. M. Blencke, G. Homuth, H. Ludwig, U. Mader, M. Hecker, and J. Stulke. Transcriptional profiling of gene expression in response to glucose in *Bacillus subtilis*: Regulation of the central metabolic pathways. *Metabolic Engineering*, 5(2):133–149, 2003.

[BSRG06] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta. BiologicalNetworks: Visualization and analysis tool for systems biology. *Nucleic Acids Research*, 34:W466–W471, 2006. Web Server Issue.

[BST03] B.-J. Breitkreutz, C. Stark, and M. Tyers. Osprey: A network visualization system. *Genome Biology*, 4(3):R22, 2003.

[Gil06] D. Gilbert. JFreeChart. http://www.jfree.org/jfreechart, 2006.

[HCL05] J. Heer, S. K. Card, and J. A. Landay. Prefuse: a toolkit for interactive information visualization. In *CHI '05: Proc. SIGCHI conf. Human factors in computing systems*, pages 421–430, 2005.

[HMM00] I. Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: a survey. *IEEE Trans. Visualization and Computer Graphics*, 6(1):24–43, 2000.

[HMW+05] Z. Hu, J. Mellor, J. Wu, T. Yamada, D. Holloway, and C. DeLisi. VisANT: Data-integrating visual framework for biological networks and modules. *Nucleic Acids Research*, 33:W352–W357, 2005. Web Server Issue.

[HMWD04] Z. Hu, J. Mellor, J. Wu, and C. DeLisi. VisANT: An online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, 5:17, 2004.

[LBKK07] A. T. Lulko, G. Buist, J. Kok, and O. P. Kuipers. Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in *Bacillus subtilis* reveals additional target genes. *Journal of Molecular Microbiology and Biotechnology*, 12(1–2):82–95, 2007.

[LCB+05] G. L. Lorca, Y. J. Chung, R. D. Barabote, W. Weyler, C. H. Schilling, and M. H. Saier Jr. Catabolite repression and activation in *Bacillus subtilis*: Dependency on CcpA, HPr, and HprK. *Journal of Bacteriology*, 187(22):7826–7839, 2005.

[LK05] W. Li and H. Kurata. A grid layout algorithm for automatic drawing of biochemical networks. *Bioinformatics*, 21(9):2036–2042, 2005.

[MNON04] Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai. DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Research*, 32:D75–77, 2004.

[MSM+01] M. S. Moreno, B. L. Schneider, R. R. Maile, W. Weyler, and M. H. Saier Jr. Catabolite repression mediated by the CcpA protein in *Bacillus subtilis*: Novel modes of regulation by whole-genome analysis. *Molecular Biology*, 39(5):1366–1381, 2001.

[SH00] J. Stulke and W. Hillen. Regulation of carbon catabolism in bacillus species. *Annual Review of Microbiology*, 54:849–880, 2000.

[SLN05] P. Saraiya, P. Lee, and C. North. Visualization of graphs with associated timeseries data. In *Proc. IEEE Symp. Information Visualization (InfoVis'05)*, pages 225–232, 2005.

[SMO+03] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003.

[The00] The Gene Ontology Consortium. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[War04] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers, 2nd edition, 2004.

[WL03] J. B. Warner and J. S. Lolkema. CcpA-dependent carbon catabolite repression in bacteria. *Microbiology and Molecular Biology Reviews*, 67(4):475–490, 2003.